LEARNING REDUCED-ORDER HAMILTONIANS FROM QUANTUM TRANSPORT: A CASE STUDY WITH DNA

Anonymous authors
Paper under double-blind review

ABSTRACT

Deep learning methods have been applied to atomic-level structure prediction and tight-binding Hamiltonian prediction, often using graph encoding of an atomic structure as an input. In this work, graph neural networks are used to construct a model Hamiltonian from an input graph that represented the interconnected bases in double-stranded DNA, with multiple cascaded layers connecting to a reduced-order model Hamiltonian. This Hamiltonian is used to calculate the quantum transmission and density of states using the non-equilibrium Green's function (NEGF) approach. The model is trained on density of states and transmission spectra collected from ab initio data from a dataset of 2,077 transmission and DOS spectra from 520 DNA sequences. Notably, no ab-initio Hamiltonian or orbital energies entered the model or training approach; only quantum transport data was used for model supervision. The Hamiltonian model was compared to a neural-network-only direct model, which pools node embeddings and feeds them into a multi-layer perceptron to construct the transmission and DOS outputs. Four architecture parameters–number of graph layers, assigned orbitals per base, supervision method, and geometry embeddings–were subjected to a full factorial analysis of variance (ANOVA) test against seven metrics to determine significance and optimize model architecture, ablating terms that did not significantly impact the model. For in-distribution data, the validation set of 4-8 base-pair (bp) duplexes, the best Hamiltonian checkpoint reaches a 1.9x higher loss than the best direct checkpoint (0.33 vs 0.17 decades of Huber loss). For out-of-distribution data, a set of four 12-bp duplexes and four 16-bp duplexes, the best Hamiltonian checkpoint beats the best direct checkpoint in each case by 0.6-2.3 decades. This chosen Hamiltonian model also reproduces the resonant transmission bands in repeating duplexes (poly-G and poly-AT), showing promise for rapid screening of other DNA variants with similar properties.

1 INTRODUCTION

The application of machine-learning techniques to electronic structure methods has been an active research area for the last decade, from decision-tree-type models Pereira et al. (2016) to deep-learning methods Choudhary et al. (2022), with force fields as the most mature application Mortazavi et al. (2023). More recently, machine learning has been used to produce the electronic structure itself, trained on ab-initio data to parameterize tight-binding models Gu et al. (2024); Wang et al. (2021), reproduce a single-electron Hamiltonian Li et al. (2022); Zhong et al. (2023); Schütt et al. (2019) or parameterize orbital energies Gu et al. (2024). While the atomic-resolution data is quite powerful for applications like materials discovery Choudhary et al. (2020), a reduced-order model can be helpful for understanding and explaining results on a larger scale, either by down-sampling a Hamiltonian or training directly on observables. For nanodevice applications in particular, models have been developed to predict quantum transport data Lopez-Bezanilla & von Lilienfeld (2014); Xie et al. (2024) or to use this data to parameterize tight-binding Hamiltonians for large nanostructures Zhouyin et al. (2023); Pawłowski & Krawczyk (2026). In this work, we go one step further: rather than incorporating the ab-initio Hamiltonian into the training data, the model is trained only on the quantum transport data, using this alone to produce a characteristic Hamiltonian of the system.

We apply this to DNA in solution, which has been shown to have non-trivial transport properties Aminiranjbar et al. (2026); Patil et al. (2021) with applications in molecular electronics and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

biosensing. DNA is a large molecule that is expensive to model with density functional theory (DFT), but its overall structure consists of only four canonical bases, which lends itself well to coarse-grained models Dans et al. (2010); Ouldridge et al. (2010). Quantum transport models have reproduced measured conductances from single-molecule-break-junction (SMBJ) experiments, but the resulting analysis shows that transport is dependent on many occupied molecular orbitals delocalized over the duplex rather than a few characteristic frontier orbitals Li et al. (2018); Qi et al. (2013). Using a reduced-order Hamiltonian can simplify this model significantly, pinpointing the exact energies associated with each base as well as the inter-base coupling parameters that are critical for charge transport.

To apply the model, DNA is represented as a graph: nodes for each base and edges for each interaction (complementary base hydrogen bond, backbone bonding, base-contact interaction). These features are embedded with identification parameters, with structural information embedded into each edge. A graph neural network (GNN) reads this structure and generates a representative Hamiltonian, and the non-equilibrium Green's function (NEGF) formalism turns this into a quantum transmission and density of states spectra that are compared against ab-initio references. In parallel, we have constructed a direct model containing only interconnected neural network layers for benchmarking, comparing the effects of pure data fitting in a neural network to our physics-informed architecture.

To our knowledge, this combination is new: learned Hamiltonians have been produced for crystals and molecules at atomic resolution Li et al. (2022); Zhong et al. (2023); Schütt et al. (2019); Unke et al. (2021), coarse-grained electronic structures have been parameterized for DNA base couplings Korol & Segal (2019); Bag et al. (2020); Aggarwal et al. (2021), and idealized Hamiltonians have been fit to transport data for solid state devices Zhouyin et al. (2023); Pawłowski & Krawczyk (2026). However, in each case the model keeps the full atomic data or is graded against the ab-initio electronic structure data itself, while our model is reduced-order and graded solely on the quantum transmission data. This data can be encoded globally, as in the full transmission through the DNA duplex, or locally, using the spectral function to calculate a local density of states (LDOS) for each base, constraining the learned Hamiltonian without including elements explicitly Burgarth & Maruyama (2009). To achieve this, we have designed a new architecture, tuning all design parameters in a full factorial experimental matrix against seven metrics. We have also released a training dataset containing 2,077 ab-initio transport records over 520 sequences, along with 8 held-out sequences for length extrapolation validation and the corresponding data generation pipeline.

2 METHODS

All runs were conducted with the developed `G3NAT` software package, which uses PyTorch Paszke et al. (2019) to construct and train all neural network architectures. A training dataset was constructed using the Gaussian software package Frisch et al. (2016) for all DFT calculations. Both the model and the dataset are available online, along with the analysis code used in the final ANOVA regression analysis.

2.1 DATASET

For this work a dataset of double-stranded DNA sequences was constructed, with randomly generated sequences between 4 and 8 base pairs (bp), as well as an additional testing dataset with four chosen 12-bp and four chosen 16-bp strands. These sequences were generated randomly, using the nucleic acid builder (`nab`) software package to build an idealized B-form double-stranded DNA sequence Macke & Case (1997). The generated coordinates were fed directly into a Gaussian input file, which was used to generate the Hamiltonian and extract the quantum transport data. These coordinates were also fed into the X3DNA-DSSR program Lu et al. (2015), which was used to generate the characteristic rigid body parameters for each base and base pair. For each duplex, four electrode configurations were chosen, attaching the electrodes directly to every atom in the corresponding bases and a constant coupling value was applied to each basis function on the atoms corresponding to the full nucleotide residue in the coordinate file. Two coupling strengths were used, 0.1 eV or 0.6 eV, based on previous setups compared to experiment Qi et al. (2013), with electrodes attached to the same single duplex (labeled "same") or across on separate strands in a duplex (labeled "cross"). Specifically, the "left" contact was always applied to the first primary base, while the "right" contact

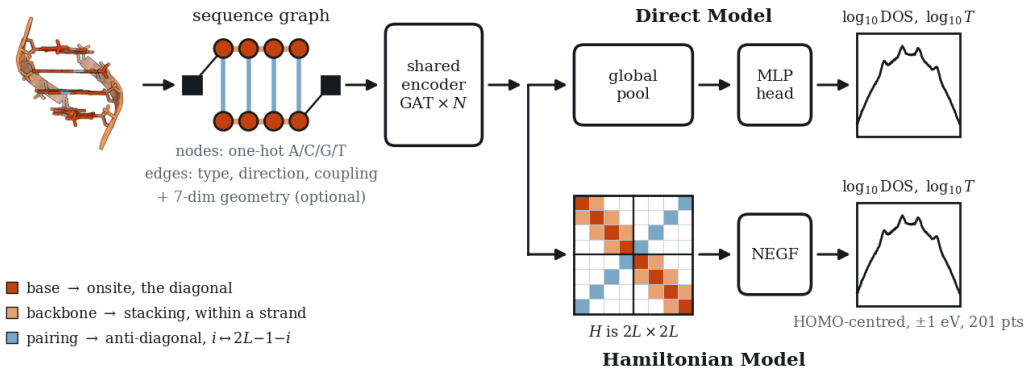

Figure 1: The overall neural network architecture employed by the G3NAT package used in this work. The direct model uses pooling and neural network models only, while the Hamiltonian model uses the NEGF formalism as a physics-informed approach. The variable L is defined as the number of basepairs in the DNA strand (1 orbital per base gives a $2L \times 2L$ matrix) and N is the number of GNN layers. All training occurs on the characteristic transmission and density of states spectra in the sampling window: $\pm$ 1 eV from the HOMO energy with a grid spacing of 0.01 eV.

was applied to the last primary base ("same") or its complement ("cross"). The dataset comprises 2,077 DFT+NEGF records over 520 duplexes of 4–8 base pairs. Each duplex was computed at four electrode configurations with the exception of two duplexes (CGTAT and GCCTGG)—the three corrupted configurations from those duplexes are absent. Only one contact configuration was used for the 12-bp and 16-bp stands (0.1 eV, cross), for a total of 8 additional NEGF records.

The DFT calculations were set up to match previous setups for quantum transport in DNA-based devices Wang et al. (2024); Patil et al. (2021). The B3LYP functional Becke (1993) was used with the 6-31G(d,p) Pople basis set for each atom type Hehre et al. (1972); Hariharan & Pople (1973), using a polarizable continuum model as an implicit water model Tomasi et al. (2005) with a net charge on the DNA equal to the number of phosphate groups corresponding to $q_{tot} = -2(N_{bp} - 1)$. From this calculation, the Löwdin transform was applied to the Fock and overlap matrices to get

$$H = S^{-\frac{1}{2}} F S^{-\frac{1}{2}} \tag{1}$$

where H is the characteristic Hamiltonian of the system in an orthogonalized basis. The highest occupied molecular orbital (HOMO) energy was used to define an energy grid, which was used for transport calculations. Applying the non-equilibrium Green's function framework Datta (2005), the characteristic retarded Green's function is given by

$$\hat{G}^r(E) = (E\hat{I} - H - \hat{\Sigma}_L^r - \hat{\Sigma}_R^r)^{-1} \tag{2}$$

a matrix output that is dependent on energy E, with $\hat{I}$ as the identity matrix and the self-energies $\hat{\Sigma}_{R/L}^r$ given by

$$\hat{\Sigma}_{R/L}^r = \text{diag}_j \begin{pmatrix} \frac{-i\Gamma_{R/L}}{2} & \text{if } j \in R/L \\ 0 & \text{if } j \notin R/L \end{pmatrix} \tag{3}$$

with $\Gamma_{R/L}$ as the contact couplings (0.1 or 0.6 eV depending on the data point) for the right/left contact and j representing the specific atomic orbital connected to the contact. The retarded Green's function can be used to construct the local density of states

$$\text{LDOS}_j(E) = -\frac{1}{\pi}\mathbf{Im}\left[\sum_{k \in j} G_{kk}^r(E)\right] \tag{4}$$

a vector with components summed across all basis functions k in base j. The total density of states is given by

$$\text{DOS}(E) = \sum_j \text{LDOS}_j(E) = -\frac{1}{\pi}\mathbf{Im}\left[\mathbf{Tr}\left[\hat{G}^r(E)\right]\right] \tag{5}$$

a scalar at each energy E. The transmission is given by

$$T(E) = \mathbf{Tr}\left[\hat{\Gamma}_L \hat{G}^r(E)\hat{\Gamma}_R(\hat{G}^r(E))^\dagger\right] \tag{6}$$

a scalar at each energy E, where $\hat{\Gamma}_{R/L}$ is a diagonal matrix with non-zero elements equal to the contact coupling at the indices corresponding to the bases that are connected to the right/left contact respectively. For this dataset, the energy grid was constructed around the HOMO energy, sampling every 0.01 eV from HOMO-1 eV to HOMO+1 eV (201 energy points total). This window was chosen for this work with the expectation that any experimental measurements would occur near this energy window. A side-effect of this choice is that all energy values are reported with respect to the highest occupied energy level rather than relative to some external vacuum or reference energy. For instance, the absolute HOMO of a GC-only system is 0.81 eV higher than an AT-only system in the DFT setup, but these would both appear as the 0 eV reference.

2.2 Model Architecture

The overall architecture of the model is laid out in Figure 1, using the structure of the double-stranded DNA duplex as an input to the GNN layers, which feed into a direct neural-network-only model and a physics-based Hamiltonian model. The GNN layer was constructed using the PyTorch Geometric Fey et al. (2025) implementation of the graph attention network (GAT) Veličković et al. (2018) message passing protocol. Input graphs use one-hot encoding for identification: nodes identify a base or contact (four one-hot dimensions for A, C, G, T, all off for contact) while edges encode connection type (three one-hot dimensions for backbone, hydrogen bond, or contact). Edges also encode directionality, coupling, and geometry, the last of which can be enabled or disabled between setups. If enabled, the geometry contains seven extracted dimensions: the center of mass distance between bases (counting the non-hydrogen base atoms, excluding sugar and backbone) and six rigid body parameters from X3DNA Lu & Olson (2003): {shift, slide, rise, tilt, roll, twist} values for nearest-neighbor bases along the backbone and {shear, stretch, stagger, buckle, propeller, opening} values for complementary (hydrogen-bonded) bases. These geometry parameters were specifically chosen because they are well established for nucleic acids and they are SE(3)-invariant, ensuring training is not rotation or translation dependent. The nodes and edges of the initial graph were projected to a single hidden layer (width 256) before entering into the GATConv layers. Both of these branches use these input GNN layers to construct output vectors of transmission and density of states (DOS) values across the specified energy grid. Specific architecture choices are summarized in Appendix A.1(Table 1), with the number of layers tested exhaustively in the following sections.

In the direct model, the GNN layers feed into a global mean pool layer, followed by projection into a multi-layer perceptron (MLP) head that feeds directly into the output vectors. This is given by

$$\mathbf{g} = \frac{1}{N}\sum_{i=1}^{N}\mathbf{h}_i, \qquad \log_{10} D(E) = \mathrm{MLP}_D(\mathbf{g}), \qquad \log_{10} T(E) = \mathrm{MLP}_T(\mathbf{g}) \tag{7}$$

where $\mathbf{h}_i$ are the node values for the N nodes of the last graph layer and MLP are the two-layer perceptrons (dimension 256 to 128 to 201 in this study). The values of these nodes are invariant to the number of nodes, and edge features are only incorporated through message-passing at each GNN layer.

In the Hamiltonian model, the nodes and edges of the last layer, omitting the contacts themselves, feed directly into values encoded into a matrix: nodes $\mathbf{h}_i$ feed into the main diagonal whereas both the nodes $\mathbf{h}_i$ and edges $\mathbf{e}_{ij}$ inform the off-diagonal terms with hermiticity enforced. This is given by the equation

$$\hat{H} = \sum_i (\hat{W}_{\mathrm{on}}(\mathbf{h}_i) + 10^{-6}\hat{I}_n)\hat{a}_i^\dagger \hat{a}_i + \sum_{i\neq j}\hat{W}_{\mathrm{cp}}([\mathbf{h}_i;\mathbf{h}_j;\mathbf{e}_{ij}])\hat{a}_i^\dagger \hat{a}_j \tag{8}$$

where n is the number of orbitals assigned to each base, $\hat{a}_i$ is the creation operator for base i, $\hat{W}_{\mathrm{on}}$ and $\hat{W}_{\mathrm{cp}}$ are $n \times n$ operators constructed from two-layer perceptrons (dimension 256 to 256 to n^2) , $\hat{I}_n$ is the identity matrix used to apply a fixed 10^{-6} eV diagonal shift for stability. Finally, the transmission and density of states values are calculated directly from this Hamiltonian. The size of this model Hamiltonian is $2nN_{\mathrm{bp}}$ orbitals for $2N_{\mathrm{bp}}$ bases in the duplex, with the size of the

reduced-order Hamiltonian scaling directly with the input graph size. That is, the model Hamiltonian is a square matrix that scales with duplex length and is mathematically connected to the output vectors, while the direct model uses pooling to determine the same output vectors. This choice was intentional to determine what features can be blindly learned from the graph architecture by itself without any physics-informed architecture.

The non-equilibrium Green's function formalism was encoded directly into the Hamiltonian model, adding a finite broadening to improve stability during initial training epochs. This modifies the construction of the Green's function in equation 2 to

$$\hat{G}^r(E) = ((E + i\eta)\hat{I} - H - \hat{\Sigma}_L^r - \hat{\Sigma}_R^r)^{-1} \tag{9}$$

where $\eta = 10^{-12}$ eV was used for this work. Note that the Green's function has the same dimensionality as the model Hamiltonian, which has dimensionality $N_b \cdot n_{orb}$ where N_b is the number of bases and n_{orb} is the number of orbitals in base b. The $\hat{\Sigma}_{R/L}^r$ matrices are constructed from equation 3 using this reduced dimensionality. This model retarded Green's function was used to calculate the density of states (equations 4 and 5) and transmission (equation 6) directly. The three outputs have dimensions that match the training set exactly: with an energy grid of 201 points, the transmission and total DOS are given as 1×201 vectors, while the LDOS is an $N_b \times 201$ matrix. The direct model has no mechanism for producing an LDOS value, and instead produces only a DOS and transmission vector output, with each grid point fully connected to the MLP head. In all cases a final `log10` is applied element-wise to each LDOS, DOS, and transmission value, with the final output vectors in decade units. This is necessary for DNA systems, where the transmission and DOS spectra can span more than eight decades in the window, and using those units drastically improves spectrum matching during training. Only a single floor was applied to every log10 function at evaluation to ensure a positive real value and guard against rounding or casting problems–this ensures that transmission and DOS values down to that limit (10^{-25}) can be used for training. All transmission and DOS training points were above this cutoff (lowest value 6.7×10^{-19}).

2.3 Training and Supervision

The Huber loss function is applied directly to the differences in the vectors in log units Huber (1964). The Adam optimizer was used for training Kingma & Ba (2017), with a learning rate of 10^{-3} and a weight decay of 10^{-5} applied. To further improve optimization and stability, linear warm-up from 10^{-5} over the first 50 epochs Goyal et al. (2017) and gradient clipping were applied during training Pascanu et al. (2013). In every setup, the total loss function was given by

$$\mathcal{L}_{tot} = a \cdot \mathcal{L}_T + c \cdot [b \cdot \mathcal{L}_{LDOS} + (1 - b) \cdot \mathcal{L}_{DOS}] \tag{10}$$

where a, b, and c were variables that were used to select the supervision method for that setup. The direct model architecture does not include the LDOS output ($b = 0$ effectively for all direct model setups), as it is impossible to decouple from the DOS without the embedded NEGF machinery. Recent work has demonstrated that including a localized loss term can significantly improve training, especially within the density functional framework von Strachwitz et al. (2026), and this was compared directly for the Hamiltonian model. For each setup, training and validation datasets were separated by sequence string, 80/20 split between training and validation, with the four contact variants of a sequence never straddling the split. Length-bucketed batches were used for training due to the size-dependent nature of the Hamiltonian model.

Each setup was trained for 10,000 epochs total, selecting the run with the lowest validation loss based on its own validation objective from equation 10. That is to say that the final model for a run trained only on the transmission loss ($a = 1, b = 0, c = 0$) is only chosen based on the run that minimizes that specific loss. This allowed us to directly compare different setups and exclude any overfitting that may have occurred after the validation objective was reached. To ensure reproducibility and test for initialization variance, three initialization seeds were chosen to create three replicate runs for each setup–while the results may not be bit-reproducible they are statistically reproducible from that seed.

2.4 Full-Factorial ANOVA Design

To fully characterize this architecture and approach, several of the design variables were iterated in a full factorial design of experiments (DoE) analysis, measuring the statistical significance of each

variable and optimal setup against seven metrics. Four variables were chosen for the analysis of the Hamiltonian model: supervision (loss function), assigned orbitals per base, number of GNN layers, and addition of geometry parameters. Table 2 in Appendix A.2 exhaustively lists all of the setups variables used, replicated across the three initialization seeds for 96 total runs: (DOS+T, LDOS+DOS+T, LDOS+T, T-only) $\times$ n_{orb} (1, 2) $\times$ depth (2, 4 layers) $\times$ geometry (off, on) $\times$ 3 seeds. A subset of these parameters was applied to the direct model, which does not have an assigned orbital variable and only has two possible supervision methods, leading to 24 total runs: supervision (DOS+T, T-only) $\times$ depth (2,4)$\times$ geometry (off, on) $\times$ 3 seeds. In both cases, the number of GNN layers was chosen as a variable based on previous results with GAT message passing Xu et al. (2019); Alon & Yahav (2021) and with 2 versus 4 layers as the chosen setpoints. These results were benchmarked against seven response metrics (also listed in Table 2), the first five of which were shared between the Hamiltonian and direct models: in-distribution transmission loss (from the validation data), transmission/DOS loss for four held-out 12-bp duplexes, transmission/DOS loss for four held-out 16-bp duplexes (held-out sequences given in Figure 4). We classify the held-out 12-bp and 16-bp duplexes as *out of distribution* because their lengths are larger than the training set of 4- to 8-bp duplexes, which are *in distribution*. The other two response metrics only applied to the Hamiltonian model, quantifying the physicality of the orbital energies produced: the minimum onsite energy distance from HOMO, averaged across all bases of all 2077 records, and onsite energy contact invariance, calculated between each of the four contact setups as the RMS energy shift of all orbitals assigned to the interior bases of the duplex not attached to the contacts (median value across the 518 duplexes with complete contact setups).

Each response was fitted with a saturated, sum-to-zero linear model over the four factors (15 terms, 64 residual degrees of freedom for the Hamiltonian family; 7 terms and 16 for the direct family), and every term was tested with a Type III drop-one F-test Speed et al. (1978), with the p-values of each family corrected together by Benjamini–Hochberg (BH) at $q = 0.05$ Benjamini & Hochberg (1995). The model, the test statistic and the correction are written out in Appendix A.2, where Tables 3 and 4 give the best configuration for every retained term.

3 Results

3.1 ANOVA Results

In lieu of a smaller subset of ablation runs, the full factorial training setup allowed us to pinpoint exactly which parameters were statistically important for achieving the five fitting metrics and two physicality metrics. The majority of the statistically significant parameters come from the Hamiltonian model, which was tested against all seven metrics, while only direct model terms that survived the BH procedure show that DOS supervision is statistically significant for the DOS extrapolation (a tautology), two of the five tested metrics. A full table of the analysis results for each metric and statistically significant terms is in Appendix A.2. Below are summaries of the major results:

3.1.1 Supervision

In this work, the supervision method was set by the $\{a, b, c\}$ variables used in equation 10 to train the model. Supervision was found to be statistically significant across six out of seven metrics, only absent on L=12 transmission. Within the $n_{orb} = 2$ family, the inclusion of the LDOS term optimized all transmission metrics (both within the dataset and extrapolating to 12-bp and 16-bp duplexes). However, the LDOS term did not improve the onsite energy locality over DOS+T supervision, as shown in Figure 2, with the nearest orbital HOMO distance slightly higher than for DOS+T supervision across all of the $n_{orb} = 2$ models.

3.1.2 Base Orbital Assignment

Assigning two orbitals per base was optimal and significant as a main effect on in-distribution (p=2.2e-15), 12-bp (p=5.2e-5), and 16-bp (p=1.8e-10) transmission loss as well as reducing the minimum onsite energy distance (p=2.2e-9). To reduce contact invariance, one orbital was preferred, and only significantly as a cross term with supervision (p=2.4e-7). This tracks as a generic statement about the inverse problem: the reduction in Hamiltonian size makes the fitting harder to achieve and causes the resulting Hamiltonian to be less physically accurate. It would be ideal to

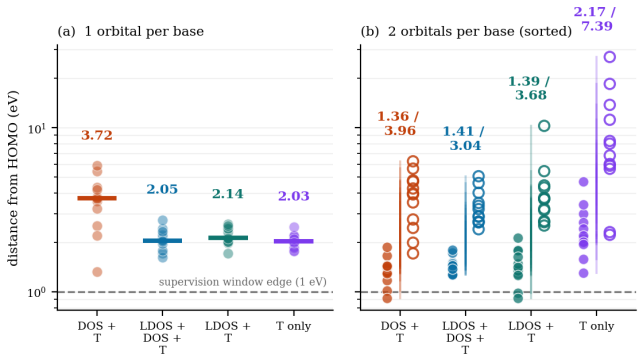

Figure 2: A comparison of the average onsite energy distance from HOMO for (a) a single orbital per base or (b) two orbitals per base, with the dataset energy grid boundary shown as a grey dotted line at 1 eV

construct a single-orbital tight-binding model that could capture the majority of the effects, but the generated transmission peaks cannot match ab-initio data and the onsite energies were found to exist mostly outside of the energy window as shown in Figure 2(a). Instead, adding a second orbital per base allowed the onsite energies to split into one "deep-band" energy and one "local" energy, shown in Figure 2(b), allowing for better fitting with transmission data constructed from thousands of interacting orbitals.

3.1.3 GNN Layers

Runs with four GNN layers, allowing message passing to four nearest neighbors, were found to be optimal and statistically significant in the Hamiltonian model for in-distribution transmission loss (p=1.9e-15) as well as in the held-out 12-bp duplexes transmission loss (p=3.9e-7) and in multiple cross terms for the 16-bp duplexes (two-way with $n_{orb} = 2$ p=7.8e-3). Contact invariance is the only response in which two layers was optimal and significant (p = 4.0e-5). Deeper message passing lets the contact nodes reach more interior sites, so a change of attachment (same vs cross strand) moves interior onsite energies by a median 0.48 eV with four layers versus 0.24 eV with two layers (median interior drift is 0.39 eV over all 518 duplexes). In the direct model, two GNN layers reduced the transmission loss across all transmissions, though this effect never passed BH screening. This shows that the Hamiltonian fit was optimized for transmission prediction with higher message passing, at the cost of contact variability from message passing, whereas the direct model was optimized with fewer message passing layers.

3.1.4 Geometry

The geometry edge embedding switch was not found to be a statistically significant main effect across the direct or Hamiltonian models, with mixed results across each metric. Geometry only shows significance through interactions with the orbital count, each term preferring no geometry embeddings. This can be explained by the dataset itself: the textbook B-DNA structure used in this work has many parameters that are not base-dependent, such as rise (standard deviation of 0.005 $\AA$ in the dataset) or propeller angle (standard deviation of 0.010 degrees in the dataset). While other parameters did experience some variation in the dataset (COM distance had a standard deviation of 0.3 $\AA$ and twist had a standard deviation of 1.01 degrees between neighboring bases), the subspace of possible geometries was limited enough that any trained weights could lead to overfitting to the data. Therefore, for this work, we have ablated these extra parameters in the final model selection.

3.2 Final Model Selection

Based on the ANOVA results, the final representative Hamiltonian model was supervised with LDOS and transmission only ($a = b = c = 1$) with four GNN layers and the geometry embeddings turned off. Two orbitals per base were chosen to improve the losses (both interpolation and extrapolation of transmission) and ensure that learned onsite energies fell within the energy window of the training

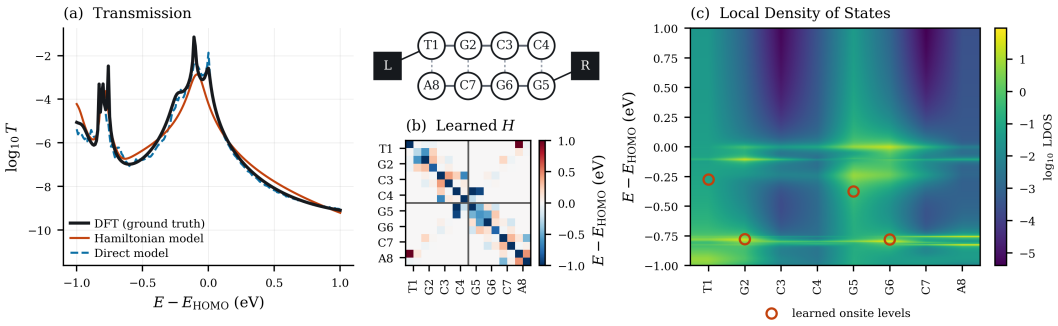

Figure 3: A demonstration of the trained Hamiltonian model on a 4-bp duplex in the dataset (TGCC, 0.1eV coupling across strands) showing (a) the comparison with the DFT ground truth and direct model, (b) the learned Hamiltonian (each base represented by 2×2 matrix for `nOrb=2`), and (c) the LDOS from the DFT ground truth data and learned onsite energies.

data. For comparison, the optimal parameters were chosen for the direct model to minimize transmission extrapolation loss, which used DOS and transmission supervision ($a = c = 1, b = 0$) with only two GNN layers and geometry embeddings turned off. Finally, from both of these models, the seed with the best overall transmission loss was chosen, scored as the sum of the three transmission loss metrics (validation, L=12, and L=16), which happened to be the same seed (3731635825) for both models. This chosen seed, both applied to a 4-base-pair duplex from the dataset (TGCC) are compared in Figure 3. As a final capacity test, two models with added parameters were compared using the same chosen seed: a *wider GNN* case with a larger hidden dimension in the GNN layer (increasing from 256 to 512) and a *deeper readout* case adding one with an additional MLP layer to the readouts in each model. The setup and results for these tests are described in detail in Appendix A.3. In both cases, the resulting losses were not significantly improved: the in-distribution transmission losses changed by less than 0.1 decade across each of the four models, with the out of distribution losses transmission losses increasing up to 0.1 decades for the direct model (12-bp duplexes with added MLP layer) and up to 1.175 decades for the transmission model (12-bp duplexes with larger GNN hidden layer). This result shows that the chosen models do not benefit from added parameters.

4 Discussion and Limitations

The direct model predicts the peaks of the transmission spectra more accurately than the Hamiltonian model across all runs, with in-distribution losses that are 1.7x lower (0.27 vs 0.46 decades of Huber loss, averaged over all configurations). This is demonstrated on the TGCC duplex shown in Figure 3: the direct model fits the transmission data more closely, even including the small HOMO peak that the Hamiltonian model cannot reproduce. This peak can be seen on the G6 base in Figure 3(c) localized at 0 eV, a location without an assigned onsite energy from the Hamiltonian model. These sharper peaks cannot be created with the reduced-order Hamiltonian, and yet the direct model is able to produce them from a coarse-grain representation of the DNA purely from message passing in the GNN layers. For out-of-distribution (12- and 16-bp) data, the transmission losses of the best direct model were higher than the best Hamiltonian model for each strand, as shown in Figure 4(c) and (d) for the 12-bp and 16-bp duplexes respectively. However, looking at the generated transmission graphs in Figure 4(a) and (b), one can see that the direct model captures the peaks better than the Hamiltonian model, especially for the heterogeneous strand in Figure 4(b). Only the off-peaks value differ, which contribute more to the overall loss value and are strictly length dependent. Still, these losses remain above those of the Hamiltonian model in the held out cases, highlighting the need for a length dependent term in the loss function for better extrapolation to the held out duplexes. While not ideal for transport prediction, the Hamiltonian model did successfully create a physically accurate reduced order Hamiltonian for the DNA duplexes. Rather than constructing fixed base-dependent energies, like a tight-binding model, message passing in the GNN allowed onsite energies and couplings to be impacted by electronic effects between neighboring bases. This also led to significant contact dependence, with a median fluctuation of 0.39eV across all inner bases between

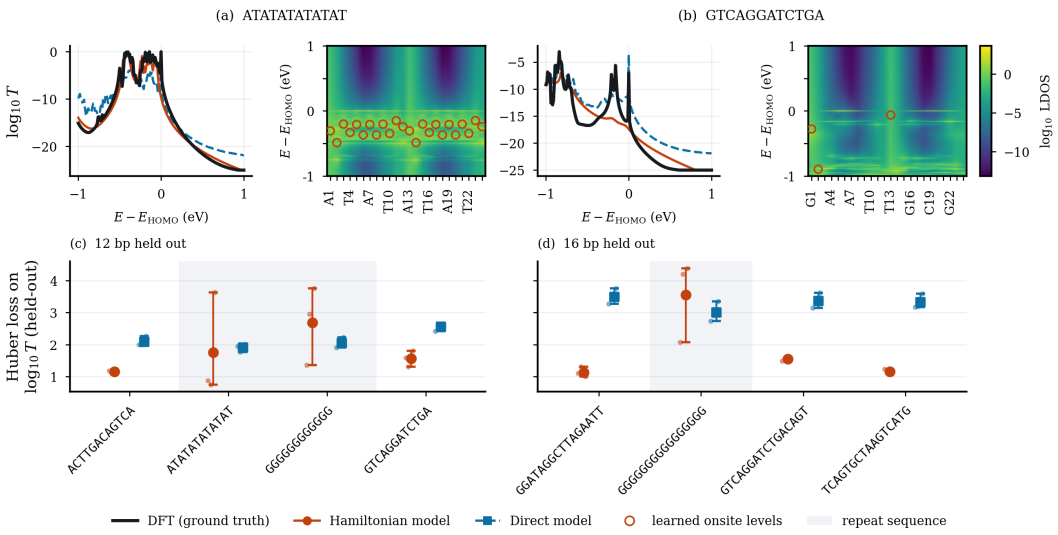

Figure 4: The model results applied to the held-out data: transmission comparison between the ab-initio data, Hamiltonian model, and direct model followed by the ab-initio LDOS data overlaid with learned onsite energies for (a) the $(AT)_6$ polymerized duplex and (b) the GTCAGGATCTGA heterogeneous duplex followed by a comparison of the $\log_{10} T$ loss for the direct and Hamiltonian models for each (c) 12-bp and (d) 16-bp duplex.

the four contact configurations. This can be a feature, showing the perturbative impact of a contact on the density of states in the bases, but can also be eliminated by excluding the contacts as graph nodes.

In summary, the Hamiltonian model was able to fit to broader energy bands, such as those with repeating bases, features that were physically possible to construct the reduced-order Hamiltonian. At these sites, the generated model Hamiltonian contains physically accurate energies, with LDOS supervision used to constrain the learned onsite and hopping energies. In Figure 3(c) and Figure 4(a) the learned energies appear exactly on these broadened peaks, whereas Figure 4(b) shows learned onsite energies only at the broadened contacts (the sharp peaks of the LDOS could not be reproduced from a reduced order Hamiltonian). This is a limitation, but only under certain conduction regimes: the transmission values associated with long heterogeneous duplexes are vanishingly small over most of the window ($T(E) < 10^{-10}$ away from the HOMO peak), whereas the resonant peaks in the polymerized AT duplex Figure 4 (a) are significant ($T(E) \sim 1$), which is within the detectable ranges of experimental SMBJ measurements. This suggests that the model architecture would be most applicable to DNA variants with high conductivity, like metal-mediated DNA systems, which contain broad resonances on each metal site Vecchioni et al. (2023) captured well by our model.

Future work could improve these initial models, starting with the supervision method. Rather than using a fixed energy grid, the intermediate basis can be changed to something more representative of the system. For instance, each resonance level contributes via a Lorentzian function Lambert (2015), which can be a basis for these output vectors. Additionally, using current-voltage characteristics generated by the NEGF layer on the model Hamiltonian, rather than transmission data under equilibrium conditions, would create new data for fitting and allow for direct comparison to experimental data. Changes can also be applied to the dataset itself to improve the model training. Experimental research shows that conductance can be modulated significantly during conformational changes Artés et al. (2015). Including multiple conformations for each sequence would not only improve training, but would also allow the trained model to be applied across a molecular dynamics trajectory, and the reduced-order Hamiltonian would indicate exactly how much onsite energies and couplings fluctuate as a function of time. This would also allow the inclusion of mismatch bases in the dataset, which would not follow B-form ordering, showing the effect of single-point mutations on the orbital energies and transmission of the system.

AI USE STATEMENT

All code for this work was created in collaboration with an LLM coding assistant (Claude Code, using Anthropic's Claude Haiku 4.5, Sonnet 4.5-5, Opus 4.5-5.5, and Fable 5-5.1 models). Feature implementation in the model and analysis scripts used test-driven development via the Superpowers agentic framework (github.com/obra/superpowers), with all code diffs and tests reviewed by the first author. Literature searches were verified and augmented using LLM agents, with all citations resolved against Crossref and arXiv. Every scientific decision was made by humans, using LLM tools to extract and summarize outcomes from the code output. All paper text was synthesized purely by the authors, using LLM tools to give feedback and corrections line by line. This feedback was reviewed by humans and applied when correct or relevant. The first author takes responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work was a computational study of DNA charge transport; no human or animal data was used and no dual-use concern was identified. The experiments reported here used approximately 430 GPU-hours, and approximately 1,500 GPU-hours including development runs. This excludes compute used by the hosted LLM coding assistant, which we cannot measure.

REPRODUCIBILITY STATEMENT

We have ensured reproducibility of the data by making the following available:

- All code used for the model construction and training is in the repository: `https://anonymous.4open.science/r/G3NAT/`.
  - All analysis scripts that produce every reported number in the manuscript are in the `analysis/` sub-directory of the repo
- All trained model checkpoints use for this work are available in a dataset: `https://doi.org/10.5281/zenodo.22964131`.
  - The final two chosen models are in the `trained_models/` sub-directory of the repository
- The DNA dataset, including the Fock and overlap matrices with a per-orbital map: `https://doi.org/10.5281/zenodo.22964053`.

Additionally, to reconstruct the results in this paper, we have used numerical seeds for the initialization of the training weights (seeds: 1179027592, 2129768291, 3731635825) and for splitting the dataset into training and validation (seed: 42).

AUTHOR CONTRIBUTIONS

[First Author] - Model Construction and Analysis, Writing/Editing

[Second Author] - Dataset Generation

[Third Author] - Writing/Editing, Supervision

[Fourth Author] - Writing/Editing, Supervision

ACKNOWLEDGMENTS

REFERENCES

D. Adams. *The Hitchhiker's Guide to the Galaxy*. Hitch-hiker's guide to the galaxy. Pocket Books, 1981. ISBN 978-0-671-66496-1. URL `https://books.google.com/books?id=SQkU8EnJUtAC`.

Abhishek Aggarwal, Vinayak Vinayak, Saientan Bag, Chiranjib Bhattacharyya, Umesh V. Waghmare, and Prabal K. Maiti. Predicting the DNA Conductance Using a Deep Feedforward Neural

Network Model. *Journal of Chemical Information and Modeling*, 61(1):106–114, 2021. doi: 10.1021/acs.jcim.0c01072.

Uri Alon and Eran Yahav. On the Bottleneck of Graph Neural Networks and its Practical Implications. In *International Conference on Learning Representations (ICLR)*, 2021.

Zahra Aminiranjbar, Caglanaz Akin Gultakti, Amy Zhang, Ersin Emre Oren, and Joshua Hihath. Developing design guidelines for controlling charge transport in DNA. *Nature Chemistry*, 18(3): 519–525, March 2026. ISSN 1755-4349. doi: 10.1038/s41557-025-01999-2. URL https://www.nature.com/articles/s41557-025-01999-2.

Juan Manuel Artés, Yuanhui Li, Jianqing Qi, M. P. Anantram, and Joshua Hihath. Conformational gating of DNA conductance. *Nature Communications*, 6(1):8870, December 2015. ISSN 2041-1723. doi: 10.1038/ncomms9870. URL https://www.nature.com/articles/ncomms9870.

Saientan Bag, Abhishek Aggarwal, and Prabal K. Maiti. Machine Learning Prediction of Electronic Coupling between the Guanine Bases of DNA. *The Journal of Physical Chemistry A*, 124(38): 7658–7664, 2020. doi: 10.1021/acs.jpca.0c04368.

Axel D. Becke. Density-functional thermochemistry. III. The role of exact exchange. *The Journal of Chemical Physics*, 98(7):5648–5652, 1993. doi: 10.1063/1.464913.

Yoav Benjamini and Yosef Hochberg. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 57(1):289–300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Daniel Burgarth and Koji Maruyama. Indirect Hamiltonian identification through a small gateway. *New Journal of Physics*, 11(10):103019, 2009. doi: 10.1088/1367-2630/11/10/103019.

Kamal Choudhary, Kevin F. Garrity, Andrew C. E. Reid, Brian DeCost, Adam J. Biacchi, Angela R. Hight Walker, Zachary Trautt, Jason Hattrick-Simpers, A. Gilad Kusne, Andrea Centrone, Albert Davydov, Jie Jiang, Ruth Pachter, Gowoon Cheon, Evan Reed, Ankit Agrawal, Xiaofeng Qian, Vinit Sharma, Houlong Zhuang, Sergei V. Kalinin, Bobby G. Sumpter, Ghanshyam Pilania, Pinar Acar, Subhasish Mandal, Kristjan Haule, David Vanderbilt, Karin Rabe, and Francesca Tavazza. The joint automated repository for various integrated simulations (JARVIS) for data-driven materials design. *npj Computational Materials*, 6(1), November 2020. ISSN 2057-3960. doi: 10.1038/s41524-020-00440-1. URL https://www.nature.com/articles/s41524-020-00440-1.

Kamal Choudhary, Brian DeCost, Chi Chen, Anubhav Jain, Francesca Tavazza, Ryan Cohn, Cheol Woo Park, Alok Choudhary, Ankit Agrawal, Simon J. L. Billinge, Elizabeth Holm, Shyue Ping Ong, and Chris Wolverton. Recent advances and applications of deep learning methods in materials science. *npj Computational Materials*, 8(1):59, April 2022. ISSN 2057-3960. doi: 10.1038/s41524-022-00734-6. URL https://www.nature.com/articles/s41524-022-00734-6.

Pablo D. Dans, Ari Zeida, Matías R. Machado, and Sergio Pantano. A Coarse Grained Model for Atomic-Detailed DNA Simulations with Explicit Electrostatics. *Journal of Chemical Theory and Computation*, 6(5):1711–1725, April 2010. ISSN 1549-9618. doi: 10.1021/ct900653p. URL https://doi.org/10.1021/ct900653p.

Supriyo Datta. *Quantum Transport: Atom to Transistor*. Cambridge University Press, 2005. doi: 10.1017/CBO9781139164313.

Matthias Fey, Jinu Sunil, Akihiro Nitta, Rishi Puri, Manan Shah, Bla z Stojanovi c, Ramona Bendias, Alexandria Barghi, Vid Kocijan, Zecheng Zhang, Xinwei He, Jan Eric Lenssen, and Jure Leskovec. PyG 2.0: Scalable Learning on Real World Graphs. *Temporal Graph Learning Workshop @ KDD*, 2025.

M J Frisch, G W Trucks, H B Schlegel, G E Scuseria, M A Robb, J R Cheeseman, G Scalmani, V Barone, G A Petersson, H Nakatsuji, X Li, M Caricato, A V Marenich, J Bloino, B G Janesko,

R Gomperts, B Mennucci, H P Hratchian, J V Ortiz, A F Izmaylov, J L Sonnenberg, D Williams-Young, F Ding, F Lipparini, F Egidi, J Goings, B Peng, A Petrone, T Henderson, D Ranasinghe, V G Zakrzewski, J Gao, N Rega, G Zheng, W Liang, M Hada, M Ehara, K Toyota, R Fukuda, J Hasegawa, M Ishida, T Nakajima, Y Honda, O Kitao, H Nakai, T Vreven, K Throssell, J A Montgomery Jr., J E Peralta, F Ogliaro, M J Bearpark, J J Heyd, E N Brothers, K N Kudin, V N Staroverov, T A Keith, R Kobayashi, J Normand, K Raghavachari, A P Rendell, J C Burant, S S Iyengar, J Tomasi, M Cossi, J M Millam, M Klene, C Adamo, R Cammi, J W Ochterski, R L Martin, K Morokuma, O Farkas, J B Foresman, and D J Fox. Gaussian 16 Revision C.01, 2016.

Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He. Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, 2017.

Qiangqiang Gu, Zhanghao Zhouyin, Shishir Kumar Pandey, Peng Zhang, Linfeng Zhang, and Weinan E. Deep learning tight-binding approach for large-scale electronic simulations at finite temperatures with ab initio accuracy. *Nature Communications*, 15(1):6772, August 2024. ISSN 2041-1723. doi: 10.1038/s41467-024-51006-4. URL https://www.nature.com/articles/s41467-024-51006-4.

P. C. Hariharan and J. A. Pople. The influence of polarization functions on molecular orbital hydrogenation energies. *Theoretica Chimica Acta*, 28(3):213–222, 1973. doi: 10.1007/BF00533485.

W. J. Hehre, R. Ditchfield, and J. A. Pople. Self—Consistent Molecular Orbital Methods. XII. Further Extensions of Gaussian—Type Basis Sets for Use in Molecular Orbital Studies of Organic Molecules. *The Journal of Chemical Physics*, 56(5):2257–2261, 1972. doi: 10.1063/1.1677527.

Peter J. Huber. Robust Estimation of a Location Parameter. *The Annals of Mathematical Statistics*, 35(1):73–101, March 1964. ISSN 0003-4851, 2168-8990. doi: 10.1214/aoms/1177703732. URL https://projecteuclid.org/journals/annals-of-mathematical-statistics/volume-35/issue-1/Robust-Estimation-of-a-Location-Parameter/10.1214/aoms/1177703732.full.

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs.LG].

Roman Korol and Dvira Segal. Machine Learning Prediction of DNA Charge Transport. *The Journal of Physical Chemistry B*, 123(13):2801–2811, 2019. doi: 10.1021/acs.jpcb.8b12557.

C. J. Lambert. Basic concepts of quantum interference and electron transport in single-molecule electronics. *Chemical Society Reviews*, 44(4):875–888, February 2015. ISSN 0306-0012. doi: 10.1039/c4cs00203b. URL https://doi.org/10.1039/c4cs00203b.

He Li, Zun Wang, Nianlong Zou, Meng Ye, Runzhang Xu, Xiaoxun Gong, Wenhui Duan, and Yong Xu. Deep-learning density functional theory Hamiltonian for efficient ab initio electronic-structure calculation. *Nature Computational Science*, 2(6):367–377, June 2022. ISSN 2662-8457. doi: 10.1038/s43588-022-00265-6.

Yuanhui Li, Juan M. Artés, Busra Demir, Sumeyye Gokce, Hashem M. Mohammad, Mashari Alangari, M. P. Anantram, Ersin Emre Oren, and Joshua Hihath. Detection and identification of genetic material via single-molecule conductance. *Nature Nanotechnology*, 13(12):1167–1173, December 2018. ISSN 1748-3395. doi: 10.1038/s41565-018-0285-x. URL https://www.nature.com/articles/s41565-018-0285-x.

Alejandro Lopez-Bezanilla and O. Anatole von Lilienfeld. Modeling electronic quantum transport with machine learning. *Physical Review B*, 89(23):235411, 2014. doi: 10.1103/PhysRevB.89.235411.

Xiang-Jun Lu, Harmen J. Bussemaker, and Wilma K. Olson. DSSR: an integrated software tool for dissecting the spatial structure of RNA. *Nucleic Acids Research*, pp. gkv716, 2015. doi: 10.1093/nar/gkv716.

Xiang-Jun Lu and Wilma K. Olson. 3DNA: a software package for the analysis, rebuilding and visualization of three-dimensional nucleic acid structures. *Nucleic Acids Research*, 31(17):5108–5121, September 2003. ISSN 0305-1048. doi: 10.1093/nar/gkg680. URL https://doi.org/10.1093/nar/gkg680.

Thomas J. Macke and David A. Case. Modeling Unusual Nucleic Acid Structures. In *ACS Symposium Series*, pp. 379–393. American Chemical Society, 1997. doi: 10.1021/bk-1998-0682.ch024.

Bohayra Mortazavi, Xiaoying Zhuang, Timon Rabczuk, and Alexander V. Shapeev. Atomistic modeling of the mechanical properties: the rise of machine learning interatomic potentials. *Materials Horizons*, 10(6):1956–1968, June 2023. ISSN 2051-6347. doi: 10.1039/d3mh00125c. URL https://doi.org/10.1039/d3mh00125c.

Thomas E. Ouldridge, Ard A. Louis, and Jonathan P. K. Doye. DNA Nanotweezers Studied with a Coarse-Grained Model of DNA. *Physical Review Letters*, 104(17):178101, April 2010. doi: 10.1103/PhysRevLett.104.178101. URL https://link.aps.org/doi/10.1103/PhysRevLett.104.178101.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training Recurrent Neural Networks, 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: an imperative style, high-performance deep learning library. In *Proceedings of the 33rd International Conference on Neural Information Processing Systems*. Curran Associates Inc., Red Hook, NY, USA, 2019.

Sunil R. Patil, Hashem Mohammad, Vivek Chawda, Niraj Sinha, Reman Kumar Singh, Jianqing Qi, and M P. Anantram. Quantum Transport in DNA Heterostructures: Implications for Nanoelectronics. *ACS Applied Nano Materials*, 4(10):10029–10037, October 2021. ISSN 2574-0970. doi: 10.1021/acsanm.1c01087. URL https://doi.org/10.1021/acsanm.1c01087.

Jarosław Pawłowski and Mateusz Krawczyk. Learning Hamiltonians for solid-state quantum simulators, March 2026. URL http://arxiv.org/abs/2603.02889. arXiv:2603.02889 [cond-mat.mes-hall].

Florbela Pereira, Kaixia Xiao, Diogo A. R. S. Latino, Chengcheng Wu, Qingyou Zhang, and Joao Aires-de Sousa. Machine Learning Methods to Predict Density Functional Theory B3LYP Energies of HOMO and LUMO Orbitals. *Journal of Chemical Information and Modeling*, 57(1):11–21, December 2016. ISSN 1549-9596. doi: 10.1021/acs.jcim.6b00340. URL https://doi.org/10.1021/acs.jcim.6b00340.

Jianqing Qi, Neranjan Edirisinghe, M. Golam Rabbani, and M. P. Anantram. Unified model for conductance through DNA with the Landauer-Büttiker formalism. *Physical Review B*, 87(8):085404, February 2013. ISSN 1098-0121. doi: 10.1103/PhysRevB.87.085404.

K. T. Schütt, M. Gastegger, A. Tkatchenko, K.-R. Müller, and R. J. Maurer. Unifying machine learning and quantum chemistry with a deep neural network for molecular wavefunctions. *Nature Communications*, 10, 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-12875-2.

F. M. Speed, R. R. Hocking, and O. P. Hackney. Methods of Analysis of Linear Models with Unbalanced Data. *Journal of the American Statistical Association*, 73(361):105–112, 1978. doi: 10.1080/01621459.1978.10480012.

Jacopo Tomasi, Benedetta Mennucci, and Roberto Cammi. Quantum Mechanical Continuum Solvation Models. *Chemical Reviews*, 105(8):2999–3094, August 2005. ISSN 0009-2665. doi: 10.1021/cr9904009. URL https://doi.org/10.1021/cr9904009.

Oliver T. Unke, Mihail Bogojeski, Michael Gastegger, Mario Geiger, Tess Smidt, and Klaus-Robert Müller. SE(3)-equivariant prediction of molecular wavefunctions and electronic densities. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Simon Vecchioni, Brandon Lu, William Livernois, Yoel P. Ohayon, Jesse B. Yoder, Chu-Fan Yang, Karol Woloszyn, William Bernfeld, M. P. Anantram, James W. Canary, Wayne A. Hendrickson, Lynn J. Rothschild, Chengde Mao, Shalom J. Wind, Nadrian C. Seeman, and Ruojie Sha. Metal-Mediated DNA Nanotechnology in 3D: Structural Library by Templated Diffraction. *Advanced Materials*, 35(29):2210938, 2023. ISSN 1521-4095. doi: 10.1002/adma.202210938. URL https://onlinelibrary.wiley.com/doi/abs/10.1002/adma.202210938.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Liò, and Yoshua Bengio. Graph Attention Networks, February 2018. URL http://arxiv.org/abs/1710.10903. arXiv:1710.10903 [stat.ML].

Antonius von Strachwitz, Karim K. Alaa El-Din, Ana C. C. Dutra, and Sam M. Vinko. Data-efficient learning of exchange-correlation functionals with differentiable DFT. *Machine Learning: Science and Technology*, 7(2):025001, February 2026. ISSN 2632-2153. doi: 10.1088/2632-2153/ae3c5a.

Kexin Wang, Pu Deng, Huili Lin, Wei Sun, and Jie Shen. DNA-Based Conductors: From Materials Design to Ultra-Scaled Electronics. *Small Methods*, 9(6):2400694, 2024. doi: 10.1002/smtd.202400694.

Zifeng Wang, Shizhuo Ye, Hao Wang, Jin He, Qijun Huang, and Sheng Chang. Machine learning method for tight-binding Hamiltonian parameterization from ab-initio band structure. *npj Computational Materials*, 7(1), January 2021. ISSN 2057-3960. doi: 10.1038/s41524-020-00490-5.

Xiaoxin Xie, Zhijiang Wang, Yuchen Wang, and Fei Liu. Global Field Heterogeneous Graph Neural Networks for Accelerating Quantum Transport Calculation. In *2024 International Conference on Simulation of Semiconductor Processes and Devices (SISPAD)*, pp. 1–4. IEEE, 2024. doi: 10.1109/SISPAD62626.2024.10733057.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations (ICLR)*, 2019.

Yang Zhong, Hongyu Yu, Mao Su, Xingao Gong, and Hongjun Xiang. Transferable equivariant graph neural networks for the Hamiltonians of molecules and solids. *npj Computational Materials*, 9(1), October 2023. ISSN 2057-3960. doi: 10.1038/s41524-023-01130-4.

Zhanghao Zhouyin, Xiang Chen, Peng Zhang, Jun Wang, and Lei Wang. Automatic differentiable nonequilibrium Green's function formalism: An end-to-end differentiable quantum transport simulator. *Physical Review B*, 108(19), November 2023. ISSN 2469-9969. doi: 10.1103/PhysRevB.108.195143.

A Appendix

A.1 Hyperparameter Settings

Table 1: Settings shared by every run. Quantities varied in the factorial design are listed in Table 2.

Setting	Value
Node features	4-dim one-hot {A, C, G, T}: contact is the zero vector
Edge features	5-dim edge: type one-hot {backbone, hydrogen bond, contact}, directionality, coupling Γ in eV; *optional 7-dim geometry with a mask*
Encoder	GAT message passing, hidden width 256, 4 heads, layer normalization, no dropout; 268,032 / 533,248 parameters at 2 / 4 layers ($+67{,}840$ with geometry)
Hamiltonian readout	$n \times n$ blocks for onsite (current node $\to \hat{W}_{\mathrm{on}}$) and off-diagonal (node 1, node 2, edge $\to \hat{W}_{\mathrm{cp}}$); 263,170 / 264,712 parameters at $n = 1$ / 2
Direct readout	global mean pool $\to$ two MLP heads ($256 \to 128 \to 201$) for DOS and T; 117,650 parameters
Energy grid	201 points, HOMO ± 1 eV, referenced per sequence
Loss	Huber ($\delta = 1$) on $\log_{10}$ values over all 201 points; floor $\log_{10}(\max(x, 10^{-25})$
Optimizer	Adam, learning rate 10^{-3}, weight decay 10^{-5}; linear warm-up from learning rate 10^{-5} over 50 epochs; gradient norm clipped at 1.0
Batches and budget	32 graphs per batch, length-bucketed; 10,000 epochs; each run reported at the epoch minimizing its own validation objective
Data split	80/20 by sequence (seed 42 used each time, validation and training dataset never mixed)

A.2 Full Factorial ANOVA

The analysis of variance testing started by constructing a regression for each metric centered around the mean, given by

$$y_{ijklr} \;=\; \mu + \alpha_i + \beta_j + \gamma_k + \delta_l + (\alpha\beta)_{ij} + \cdots + (\alpha\beta\gamma\delta)_{ijkl} + \varepsilon_{ijklr} \tag{11}$$

where y_{ijklr} is the value of one response (for example the 16-bp transmission loss) for the run at supervision arm i ($i = 1, \ldots, 4$), orbitals per base j ($j = 1, 2$), depth k ($k = 2, 4$ layers), geometry setting l (off, on) and initialization seed r ($r = 1, 2, 3$); μ is the grand mean over all 96 runs; α_i, β_j, γ_k and δ_l are the main effects, the shift each setting produces on its own relative to μ; $(\alpha\beta)_{ij}$ is the two-way interaction of supervision and orbital count, the additional shift when the two settings are combined beyond their main effects, and the ellipsis stands for the remaining five two-way, four three-way and one four-way products (15 terms in all); ε_{ijklr} is the residual. Each effect is constrained to sum to zero over the levels of every factor it involves, which makes the decomposition unique. Because all 15 terms are retained, the model reproduces every one of the 32 cell means exactly and the residual is the scatter of the three seeds within a cell; its variance is the seed noise floor reported with every table, with $96 - 32 = 64$ residual degrees of freedom. The direct model uses the same form with three factors (7 terms, 8 cells, 24 runs, 16 residual degrees of freedom). The analysis includes the main effects (isolated terms) and cross terms, using a Type III drop-one F-test each time to determine the significance with F given by

$$F \;=\; \frac{(\mathrm{RSS}_{-t} - \mathrm{RSS}_{\mathrm{full}})/d_t}{\mathrm{RSS}_{\mathrm{full}}/d_{\mathrm{res}}} \tag{12}$$

where $\mathrm{RSS}_{\mathrm{full}}$ is the residual sum of squares of the full model of Eq. (11); RSS_{-t} is the residual sum of squares after refitting with only the term t removed and every other term kept; d_t is the number of degrees of freedom of that term (3 for the supervision main effect, 1 for a two-level factor, and their product for an interaction); and d_{res} is the residual degrees of freedom (64 or 16). F compares the variation only that term can explain, per degree of freedom, with the seed noise per degree of freedom; the p-value is the tail probability of F under the F-distribution with (d_t, d_{res}) degrees of freedom (a Type III drop-one test) Speed et al. (1978).

The exact parameters used for each run (including the initialization seeds) are given in Table 2 below, as well as the seven response metrics used for ANOVA analysis. All runs share one 80/20 sequence split (split seed 42, from Adams (1981)) and are evaluated at the epoch minimizing each run's own validation objective. Because 105 such tests are run on the Hamiltonian family (15 terms across 7 responses) and 35 on the direct family, the p-values of each family are corrected together with the Benjamini–Hochberg (BH) procedure Benjamini & Hochberg (1995): they are ranked from smallest to largest, and a term is retained if its p-value is at most 0.05 times its rank divided by the family size, which holds the expected fraction of false discoveries among the retained terms at 5%. The resulting thresholds are $p \leq 0.0105$ (22 of 105 terms retained) and $p \leq 0.0057$ (4 of 35). For every significant term, the best configuration are given in Tables 3 and 4 below. A direct comparison of each run organized by supervision method is given in Figure 5 across each of the seven metrics.

Table 2: The exact parameters varied between the two factorial designs, analyzed independently, with every setup run at three initialization seeds (see Appendix A.2 for seed values). Response metrics tested for each model are listed afterward. Supervision arms are the loss weights (a, b, c) of Eq. 10, $b = 0$ intrinsically for the direct model.

Factor	**Hamiltonian model**	**Direct model**
Supervision (a, b, c)	DOS+T $(1, 0, 1)$; LDOS+DOS+T $(1, \frac{1}{2}, 1)$; LDOS+T $(1, 1, 1)$; T-only $(1, 0, 0)$	DOS+T $(1, 0, 1)$; T-only $(1, 0, 0)$
Orbitals per base	1, 2	—
GNN layers	2, 4	2, 4
Geometry channel	off, on	off, on
Seeds	1179027592, 2129768291, 3731635825	
Cells × seeds = runs	$32 \times 3 = 96$	$8 \times 3 = 24$
Responses tested	7	5
Tests per family, m	105	35
Response Metrics:	Transmission loss: in-distribution (L=4–8)	Transmission loss: in-distribution (L=4–8)
	Transmission/DOS loss: L=12 (4 duplexes)	Transmission/DOS loss: L=12 (4 duplexes)
	Transmission/DOS loss: L=16 (4 duplexes)	Transmission/DOS loss: L=16 (4 duplexes)
	mean onsite distance from HOMO	*(no direct model equivalent)*
	median contact invariance	*(no direct model equivalent)*

Table 3: Hamiltonian family: every term with raw $p < 0.05$ on any response, grouped by response. Drop-one (Type III) F-tests on the saturated 15-term model, $n = 96$ runs, 64 residual degrees of freedom. **Bold** rows survive Benjamini–Hochberg at $q = 0.05$ over the family of $m = 105$ tests (threshold $p \leq 0.0105$). Every response is a loss or distance, lower is better, including onsite distance (mean |level nearest the HOMO|, eV; closer to the window) and contact drift (interior-site level shift under a contact change, eV; less drift). Choice was determined by the lowest average metric value for a single main effect or combination of crossed factors.

Response	Term	F	p	Choice (means)
in-distribution T	**layers**	**109.0**	1.9×10^{-15}	**4 layers (0.498 / 0.412)**
	$\mathbf{n}_{\mathrm{orb}}$	**108.1**	2.2×10^{-15}	**$\mathbf{n}_{\mathrm{orb}} = 2$ (0.498 / 0.413)**
	supervision	**24.2**	1.4×10^{-10}	**DOS+T (0.412 / 0.436 / 0.467 / 0.505)**
	$\mathbf{n}_{\mathrm{orb}}$ × geometry	**9.6**	2.9×10^{-3}	**$\mathbf{n}_{\mathrm{orb}} = 2$, geometry off (0.393)**
	n_{orb} × layers × geometry	4.5	3.8×10^{-2}	$n_{\mathrm{orb}} = 2$, 4 layers, geometry off (0.348)
12-bp T	**layers**	**32.0**	3.9×10^{-7}	**4 layers (2.71 / 2.34)**
	$\mathbf{n}_{\mathrm{orb}}$	**18.8**	5.2×10^{-5}	**$\mathbf{n}_{\mathrm{orb}} = 2$ (2.67 / 2.38)**
	n_{orb} × layers	4.5	3.7×10^{-2}	$n_{\mathrm{orb}} = 2$, 4 layers (2.12)
12-bp DOS	**supervision**	**48.0**	2.2×10^{-16}	**DOS+T (0.184 / 0.223 / 0.258 / 0.651)**
	supervision × n_{orb}	**6.6**	5.9×10^{-4}	**DOS+T,** $n_{\mathrm{orb}} = 2$ **(0.159)**
	supervision × n_{orb} × geometry	4.0	1.2×10^{-2}	DOS+T, $n_{\mathrm{orb}} = 2$, geometry off (0.158)
16-bp T	**$\mathbf{n}_{\mathrm{orb}}$**	**57.4**	1.8×10^{-10}	**$\mathbf{n}_{\mathrm{orb}} = 2$ (2.52 / 2.21)**
	supervision	**4.7**	5.2×10^{-3}	**DOS+T (2.25 / 2.40 / 2.45 / 2.36)**
	$\mathbf{n}_{\mathrm{orb}}$ × geometry	**8.1**	5.9×10^{-3}	**$\mathbf{n}_{\mathrm{orb}} = 2$, geometry off (2.15)**
	$\mathbf{n}_{\mathrm{orb}}$ × layers	**7.6**	7.8×10^{-3}	**$\mathbf{n}_{\mathrm{orb}} = 2$, 4 layers (2.18)**
	supervision × n_{orb} × **layers**	**4.3**	8.4×10^{-3}	**DOS+T,** $n_{\mathrm{orb}} = 2$**, 4 layers (2.05)**
16-bp DOS	**supervision**	**52.5**	3.0×10^{-17}	**DOS+T (0.185 / 0.206 / 0.233 / 0.634)**
	supervision × n_{orb}	**6.1**	1.0×10^{-3}	**DOS+T,** $n_{\mathrm{orb}} = 2$ **(0.166)**
	supervision × n_{orb} × **geometry**	**5.4**	2.2×10^{-3}	**DOS+T,** $n_{\mathrm{orb}} = 2$**, geometry off (0.165)**
	n_{orb} × geometry	4.7	3.5×10^{-2}	$n_{\mathrm{orb}} = 2$, geometry off (0.272)
onsite distance	**supervision** × n_{orb}	**22.7**	4.0×10^{-10}	**DOS+T,** $n_{\mathrm{orb}} = 2$ **(1.33)**
	$\mathbf{n}_{\mathrm{orb}}$	**48.5**	2.2×10^{-9}	**$\mathbf{n}_{\mathrm{orb}} = 2$ (2.46 / 1.67)**
	supervision	**7.9**	1.5×10^{-4}	**LDOS+DOS+T (2.44 / 1.77 / 1.83 / 2.21)**
	supervision × n_{orb} × layers	3.2	3.0×10^{-2}	DOS+T, $n_{\mathrm{orb}} = 2$, 2 layers (1.32)
contact drift	**supervision** × n_{orb}	**14.6**	2.4×10^{-7}	**T-only,** $n_{\mathrm{orb}} = 1$ **(0.214)**
	layers	**19.5**	4.0×10^{-5}	**2 layers (0.355 / 0.722)**
	supervision	**5.5**	2.0×10^{-3}	**LDOS+T (0.817 / 0.436 / 0.376 / 0.526)**
	supervision × n_{orb} × layers	3.7	1.7×10^{-2}	T-only, $n_{\mathrm{orb}} = 1$, 2 layers (0.168)
	supervision × layers	2.8	4.7×10^{-2}	T-only, 2 layers (0.241)

Table 4: Direct family: every term with raw $p < 0.05$ on any response. Drop-one (Type III) F-tests on the saturated 7-term model, $n = 24$ runs, 16 residual degrees of freedom. **Bold** rows survive Benjamini–Hochberg at $q = 0.05$ over $m = 35$ tests (threshold $p \leq 0.0057$). Lower is better on every response; no onsite or contact response exists because the direct model produces no Hamiltonian. Choice definition matches Table 3. The in-distribution T response has no term with $p < 0.05$.

Response	Term	F	p	Choice (means)
in-distribution T	*none with* $p < 0.05$	–	–	–
12-bp T	layers	7.3	1.6×10^{-2}	2 layers (2.08 / 3.52)
12-bp DOS	**supervision**	**4280.0**	7.3×10^{-21}	**DOS+T (0.112 / 0.326)**
	supervision × layers	**18.9**	5.0×10^{-4}	**DOS+T, 4 layers (0.101)**
	layers	4.9	4.2×10^{-2}	4 layers (0.222 / 0.215)
16-bp T	layers	7.8	1.3×10^{-2}	2 layers (3.21 / 5.25)
16-bp DOS	**supervision**	**1343.7**	7.3×10^{-17}	**DOS+T (0.136 / 0.301)**
	supervision × layers	**11.5**	3.7×10^{-3}	**DOS+T, 4 layers (0.126)**

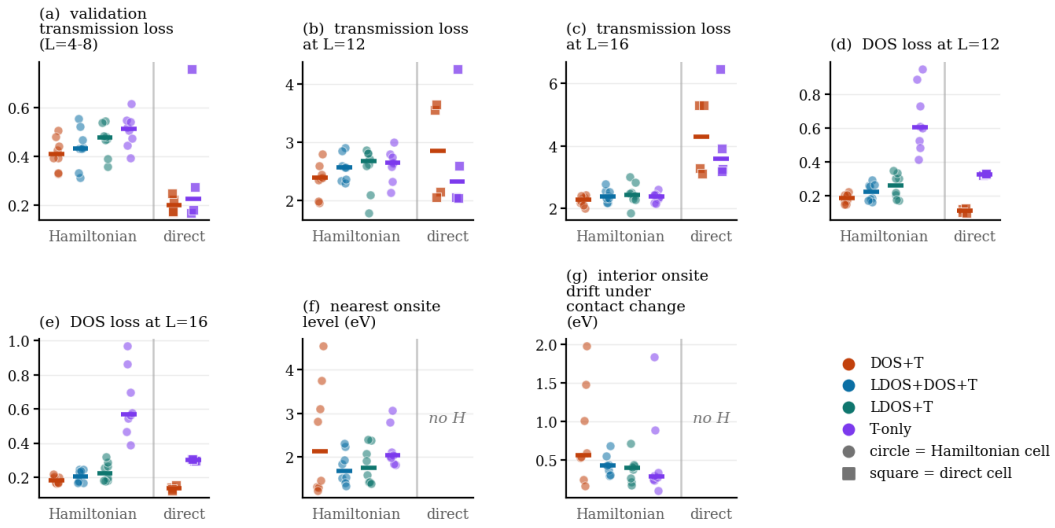

Figure 5: Every response for both families, sorted by supervision method, in the loss's own units. Each point is one configuration's cell mean over 3 seeds (8 cells per Hamiltonian arm, 4 per direct arm); the bar is the arm's median. The x-axis carries the four Hamiltonian supervision arms, a divider, then the two direct arms (squares). (a)–(c) transmission loss in-distribution and at 12-and 16-bp; (d)–(e) DOS loss at 12- and 16-bp; (f)–(g) onsite distance and contact drift, Hamiltonian only (the direct model has no H).

A.3 Capacity probe

Both reported models were retrained with two independent increases in capacity, at the seed and the recipe of the reported runs (10,000 epochs, model seed 3731635825, split seed 42), so that the effect of capacity can be separated from the effect of the readout. A description of the two models are as follows, with the tablated results below:

The *deeper readout* arm adds one $\mathrm{Linear}(256, 256)$ layer and its ReLU inside the readout perceptrons and changes nothing else: for the Hamiltonian model that is both of the perceptrons that build $\hat{H}$, $\hat{W}_{\mathrm{on}}$ $(256 \to 256 \to 256 \to n^2)$ and $\hat{W}_{\mathrm{cp}}$ $(768 \to 256 \to 256 \to n^2)$, and for the direct model both output heads $(256 \to 256 \to 128 \to 201)$. The input projections, the GAT layers and the layer normalizations are untouched, and the encoder parameter counts of Table 1 are unchanged.

The *wider GNN* arm instead sets the hidden width to 512 throughout. This widens the node and edge projections, every GAT layer (4 heads of width 128) and every layer normalization. This propagates to the readout inputs, which scale as well ($\hat{W}_{\mathrm{on}} : 512 \to 512 \to n^2$, $\hat{W}_{\mathrm{cp}} : 1536 \to 512 \to n^2$; direct heads $512 \to 256 \to 201$).

Table 5: Capacity probe: the two reported models retrained with a deeper readout or a wider GNN, at initialization seed 3731635825 and otherwise the recipe of Table 1. Every entry is a Huber loss on $\log_{10}$ values over all 201 energy points, lower is better; in-distribution T is the validation loss at the epoch the run selected, and the held-out entries pool the four duplexes of each length over their four contact configurations.

Model	**Arm**	**Parameter Count**	**Epoch**	**in-dist.** T	**12-bp** T	**16-bp** T	**12-bp DOS**	**16-bp DOS**
Hamiltonian	as reported	797,960	5516	0.332	1.245	1.471	0.150	0.165
	deeper readout	929,544	3414	0.327	2.402	2.312	0.169	0.174
	wider GNN	3,168,776	1491	0.383	2.420	2.374	0.172	0.162
Direct	as reported	385,682	728	0.171	2.068	3.079	0.129	0.155
	deeper readout	517,266	312	0.152	2.169	3.060	0.090	0.122
	wider GNN	1,426,322	293	0.167	2.019	3.163	0.097	0.117